

Probing the topological properties of complex networks modeling short written texts

Diego R. Amancio

Department of Computer Sciences
 Institute of Mathematical and Computer Sciences
 University of São Paulo, São Carlos, São Paulo, Brazil
 E-mail: diego@icmc.usp.br, diego.rafael@gmail.com

Abstract. In recent years, graph theory has been widely employed to probe several language properties. More specifically, the so-called word adjacency model has been proven useful for tackling several practical problems, especially those relying on textual stylistic analysis. The most common approach to treat texts as networks has simply considered either large pieces of texts or entire books. This approach has certainly worked well – many informative discoveries have been made this way – but it raises an uncomfortable question: could there be important topological patterns in small pieces of texts? To address this problem, the topological properties of subtexts sampled from entire books was probed. Statistical analyzes performed on a dataset comprising 50 novels revealed that most of the traditional topological measurements are stable for short subtexts. When the performance of the authorship recognition task was analyzed, it was found that a proper sampling yields a discriminability similar to the one found with full texts. Surprisingly, the support vector machine classification based on the characterization of short texts outperformed the one performed with entire books. These findings suggest that a local topological analysis of large documents might improve its global characterization. Most importantly, it was verified, as a proof of principle, that short texts can be analyzed with the methods and concepts of complex networks. As a consequence, the techniques described here can be extended in a straightforward fashion to analyze texts as time-varying complex networks.

1. Introduction

Graph theory has been employed to probe the statistical properties of many real systems [1]. Most of the real networks share the small-world [2] and scale-free [3] properties. The last fifteen years have witnessed the increase of networked models in interdisciplinary applications, including implementations in Physics [4, 5], Social Sciences [6–8], Biology [9,10], Neuroscience [11,12], Cognitive Sciences [13,14], Music [15] and Computer Sciences [16]. In the latter, graph-based techniques have been applied to the analysis and construction of software architecture [17], supervised classifiers [18], spam filters [19] and natural language processing (NLP) systems [20]. For the purpose of textual analysis, networks have proven relevant not only to improve the performance of

NLP tasks [21–23], but also to better understand the emergent patterns and mechanisms behind the origins of the language [20].

Due to its interdisciplinary nature, graph theory can be employed to study the various levels of complexity of the language. In the neuroscience field, the network framework turned out to be a powerful tool for representing the topology of neural systems, where the neocortex is responsible for conveying information [24]. Among several findings, it has been shown that the clustering and small-world effects facilitate local and global processes, respectively [14]. The connection between neuroscience and language/mind processing has been investigated in terms of the topological properties of the connectome [25]. Interestingly, it has been found that some linguistic impairments account for fluctuations in the properties of several networks representing brain organization [26–28]. At the cognitive level, networks have proven useful to unveil the mechanisms behind information processing [29]. In this context, some studies link certain diseases with specific characteristics of semantic free-association networks [30], while other investigations relate topological network properties with cognitive complexity. For example, the authors in [31] found that the recognition of a word depends on the average clustering coefficient of the network. Another example is the use of network measurements for quantifying the cognitive complexity of finding the way out of mazes [32].

Network-based models have been applied to study several levels of language organization, which encompasses both the syntactical [33, 34] and the semantical level [35–40]. A well-known model is the so-called word adjacency network, which consists in linking adjacent words. Since this model reflects mostly syntactical and stylistic factors [41], it has been successfully employed for syntactical complexity analysis [42], detection of literary movements [43] and for stylometry [44, 45]. In most applications, it has been assumed the availability of long texts (or books) to perform statistical analyses [18, 33, 41, 45–48]. Unfortunately, in many real-world situations, the availability of large pieces of texts is uncertain. Whenever only very short pieces of text are available, the conclusions drawn from the analysis could be invalid due to statistical fluctuations present in short written texts. In addition, the unavoidable noise pervading short texts could undermine the performance of NLP tasks. Therefore, it becomes of paramount importance to know beforehand if a given text is long enough for the analysis. In this context, this paper investigates how the selection of short pieces of texts (hereafter referred to as samples) affects the topological analysis of word adjacency networks.

In order to study the fluctuations of networks measurements modeling short texts, books were sampled in adjacent, non-overlapping subtexts. As I shall show, the sampling of texts yields subtexts with similar topology, as revealed by a systematic analysis of the variability of several measurements across distinct samples. The influence of the subtext length on the authorship recognition task was also studied. The results revealed that the best performance was achieved when the books were split in shorter subtexts, which confirms that the sampling might favor the classification process as the local topological

characterization of books becomes more precise.

Materials and Methods

In this section, the word adjacency model is presented. In addition, I swiftly describe the measurements employed for characterizing networks and the methods used for recognizing topological patterns.

Modeling texts as complex networks

The overall purpose of this paper is to study features of the language that reflect particular choices made by individuals or groups. Such particular choices, referred to as stylistic features, can be employed e.g. to classify genres, dialects and literary works [20]. A traditional stylistic feature used e.g. to identify authorship is the frequency of a particular function word in a text [49]. Here, a network model is used to capture particular connectivity patterns that might be useful to identify authorship, genres, languages etc.

There are several ways to model texts as complex networks [50]. The most appropriate modeling depends on the target application. One of the most employed models for grasping stylistic features is the so-called co-occurrence (or adjacency network) [41, 46]. Besides capturing syntactical attributes of the texts, networked models have also proven useful to capture language dependent features [41]. Co-occurrence networks have been employed in many applications [50]. In generic terms, a co-occurrence network can be defined in a manifold way: two words are linked if they co-occur in at least one window. The window can be represented as n-grams, sentences, paragraphs or even entire documents [51]. Some alternative co-occurrence models link two words only if the co-occurrence frequency exceeds a given threshold. There is also the possibility to improve the model by including weighted links [52]. Statistical analysis revealed that most word co-occurrence networks display both small-world and scale-free behaviors [53]. The average clustering coefficient and the average nearest neighbor degrees were found to follow a power-law distribution, as a consequence of the existence of distinct functional classes of words [54]. Particularly, it has been shown adjacency and syntactic networks display similar topological properties, as far as topological attributes are considered [33]. Besides being useful for analyzing styles in texts, co-occurrence graphs serve to model semantical relationships [55, 56].

Prior to the transformation of the text into a network model, some pre-processing steps are usually applied. Firstly, words conveying low semantic content (such as articles and prepositions) are removed. These words, referred to as *stopwords*, are disregarded from the analysis because they simply serve to connect content words. Even though previous studies used the frequency of stopwords to classify texts according to their styles, I decided not to use them because I am interested in the relationships between words with significant semantic content. This procedure has been applied in many

studies (see e.g. [44, 54, 57, 58]). In the next step, each word is transformed into its canonical form so that conjugated verbs and nouns are respectively mapped to their infinitive and singular forms. As such, distinct forms of the same word are mapped to the same concept. To obtain the canonical form of words, it is necessary to perform a sense disambiguation [51] at the word level. To assist the disambiguation process, the part-of-speech of each word is inferred from a maximum entropy model [59].

After the pre-processing step, each distinct word remaining in the text becomes a node in the network. Therefore, the total number of nodes will be given by the vocabulary size of the pre-processed text. Edges linking two words are created if these words appeared as neighbors in the text at least once. For example, in the short sentence “*Complex network measurement*”, the following links are created: *complex* \rightarrow *network* and *network* \rightarrow *measurement*. In this case, *complex* and *measurement* are not connected to each other because they are separated by one intermediary word. Particularly, this model was chosen here because it has been successfully used in applications where authors’ styles represent an important feature for analyzing written texts [53, 60]. Table 1 and Figure 1 illustrate the creation of a word adjacency network.

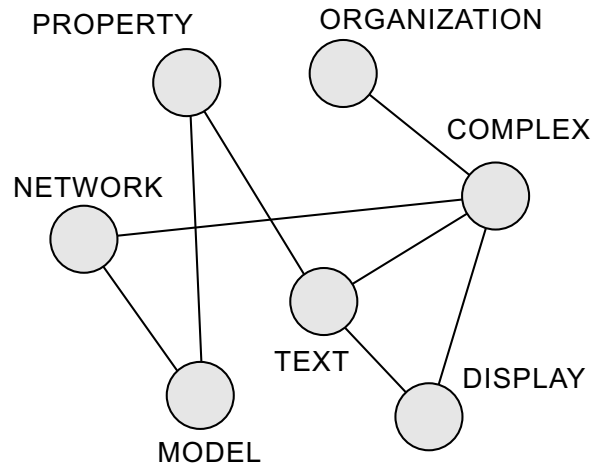


Figure 1. Example of adjacency network created from the extract: “Complex networks model several properties of texts. A complex text displays a complex organization”. After the pre-processing step, words are mapped into nodes, which in turn are connected if the respective words appeared at least once as neighbors.

Topological characterization of textual networks

The topological analysis of complex networks can be conducted through the use of a myriad of measurements. In the current study, we employed the main measurements that have been used in the analysis of word adjacency networks. The quasi-local topology, which considers the connectivity of neighbors, neighbors of neighbors and further hierarchies [61], was measured with the clustering coefficient (C) and with the

Table 1. Pre-processing steps.

Original Text	After pre-processing	Canonical form
Complex networks model	complex networks model	complex network model
several properties of	properties	property
texts. A complex text	texts complex text	text complex text
displays a complex	displays complex	display complex
organization.	organization	organization

Example of pre-processing steps performed in a extract for the purpose of creating a adjacency network. The pre-processing step eliminates punctuation marks and words conveying low semantic content. The lemmatization step aims at mapping each word to its canonical form.

average neighbor degree (k_n). Both measurements have been employed to identify words appearing in generic contexts [57]. In addition to these measurements, the accessibility (α) was used to characterize the quasi-local structure of textual networks. To define this measurement, consider that $P_{ij}^{(h)}$ represents the probability of a random walker starting at node v_i to reach node v_j in h steps. Mathematically, the accessibility $\alpha_i^{(h)}$ is defined as the entropy of the quantity $P_{ij}^{(h)}$:

$$\alpha_i^{(h)} = \exp \left(- \sum P_{ij}^{(h)} \ln P_{ij}^{(h)} \right). \quad (1)$$

It has been show that, when the walker performs a self-avoiding random walk, the accessibility is useful to identify the core of textual networks [62]. Moreover, this measurement has proven useful to generate informative extractive summaries [34].

In addition to the quasi-local measurements, we also employed some global measurements. The average shortest path length (l) was used because this measurement has been useful in several textual applications [57]. Usually, the words taking the lowest values of l are either keywords or words that appear near to the most relevant words in the text. Shortest paths were also employed in the so called betweenness (B) [63]. In textual networks, the betweenness grasps the amount of contexts in which a given word may occur [57]. Differently from the clustering coefficient, the betweenness uses the global information of the network to infer the quantity of semantic contexts in which a word occurs.

The spatial distribution of words along the text was studied in terms of the intermittency, a measurement that is able to capture the irregularity (bursts) of the distribution [64]. To compute the intermittency, one represents the pre-processed text as a time series. As a consequence, the first word is the first element of the time series, the second word is the second element and so forth. The recurrence time t_i of a word w is defined as the number of words between two successive occurrences of w . Therefore, a word occurring N times in the text generates the sequence $T = \{t_1, t_2, \dots, t_{N-1}\}$ of recurrence times. In order to consider the time t_I until the first occurrence of the word and the time t_F after the last occurrence, this measurement considers also the time

$t_N = t_I + t_F$ in T . If $\langle t \rangle$ is the mean of the elements in T , then the intermittency I of the distribution T is

$$I = \left[\frac{\langle t^2 \rangle}{\langle t \rangle^2} - 1 \right]^{1/2}. \quad (2)$$

The intermittency has been employed to identify keywords using features not correlated with word frequencies [64]. Moreover, this measurement has proven useful to classify texts according to the informativeness criterium [41].

Pattern recognition methods

In this study, we investigated if the topology of small pieces of texts is able to provide relevant attributes for textual analysis. To quantify the effects of sampling large texts in applications based upon the classification of distinct styles, supervised pattern recognition methods were used [65]. In a supervised classification problem, we are given a *training set* and a *test set*. The training set $\mathcal{S}_{tr} = \{\beta_{(tr,1)}, \beta_{(tr,2)}, \dots, \beta_{(tr,n)}\}$ is the set of examples that inference algorithms employ to generate classification models. After the creation of the classification model, the test dataset $\mathcal{S}_{ts} = \{\beta_{(ts,1)}, \beta_{(ts,2)}, \dots, \beta_{(ts,m)}\}$ is employed to evaluate the classification performance. The result of the classification is the mapping $\mathcal{S}_{ts} \mapsto \mathcal{C} = \{c_1, c_2, \dots\}$. In other words, a conventional classifier assigns a unique class $c_i \in \mathcal{C}$ for each element of the training set. For each example β , the value of the attribute F_i taken by β is represented as $\beta^{(i)}$. The performance of the classification was verified with the well known 10-fold cross-validation technique [65].

The supervised classifiers employed in this paper were: nearest neighbors (kNN) [66], decision trees (C4.5) [67], bayesian decision (Bayes) [68] and support vector machines (SVM) [69]. Below I present a swift description of these methods. Further details can be found in [65].

Nearest neighbors: this technique classifies a new example $\beta \in \mathcal{S}_{ts}$ according to a voting process performed on \mathcal{S}_{tr} . If most of the κ nearest neighbors of β belongs to the class $c_i \in \mathcal{C}$, then the class c_i is associated to β_{ts} .

Bayesian decision: this method computes the probability $P(c_i|\beta)$ that a given class $c_i \in \mathcal{C}$ is the correct class associated to a given instance $\beta_{(ts)}$. Assuming that the attributes are independent, $P(c_i|\beta)$ can be computed as

$$P(c_i|\beta) = \frac{P(c_i)}{P(F_1 = \beta^{(1)}, \dots)} \prod_k P(F_k = \beta^{(k)}|c_i). \quad (3)$$

Therefore, the correct class c_β is

$$c_\beta = \arg \max_{c_i \in \mathcal{C}} P(c_i) \prod_k P(F_k = \beta^{(k)}|c_i). \quad (4)$$

Decision trees: this algorithm is based upon the induction of a tree, a widely employed abstract data type. To construct a tree model, it is necessary to find the most informative attribute, i.e. the attribute that provides the best discriminability

of the data. To do so, several measurements have been proposed [65]. In this paper, we use the information gain Ω , which is mathematically defined as

$$\Omega(\mathcal{S}_{tr}, F_k) = \mathcal{H}(\mathcal{S}_{tr}) - \mathcal{H}(\mathcal{S}_{tr}|F_k), \quad (5)$$

where $\mathcal{H}(\mathcal{S}_{tr})$ is the entropy of the dataset \mathcal{S}_{tr} and $\mathcal{H}(\mathcal{S}_{tr}|F_k)$ is the entropy of the dataset when the value of F_k is specified. $\mathcal{H}(\mathcal{S}_{tr}|F_k)$ can be computed from the training dataset as

$$\mathcal{H}(\mathcal{S}_{tr}|F_k) = \sum_{v \in V(F_k)} \frac{|\beta_{(tr)} \in \mathcal{S}_{tr} | \beta_{(tr)}^{(k)} = v|}{|\mathcal{S}_{tr}|} \cdot \mathcal{H}(\{\beta_{(tr)} \in \mathcal{S}_{tr} | \beta_{(tr)}^{(k)} = v\}), \quad (6)$$

where $V(F_k)$ is the set of all values taken by the attribute F_k in the training dataset, i.e.

$$V(F_k) = \bigcup_{i=1}^{|\mathcal{S}_{tr}|} \beta_{(tr,i)}^{(k)}. \quad (7)$$

Support Vector Machines: this technique divides the attribute space using hyperplanes, so that each region is assigned to a single class. The construction of the hyperplanes relies upon the definition of linear or non-linear kernel functions. Once the separation is determined, a new example can be classified by evaluating its position on the attribute space. This method has been applied in several real applications due to its robustness with regard to the number of dimensions and other features [70].

Results

Variability of measurements

In this paper, a set of subtexts sampled from a entire book is considered as consistent when the measurements computed for the subtexts display low variability across different subtexts. I take the view that authors tend to keep their styles across distinct portions of the same book. This assumption is reasonable because it has been shown that the main factors responsible for stylistic variations in texts are the language [41], the authorship [57], the complexity [42] and the publication date [43]. As such, it is natural to expect low variability across distinct samples since all these factors remain constant in the same book. Hence, I consider that the main factor accounting for the variability of the style across distinct parts of the same book is the sample size.

In order to compute the variability of the measurements across distinct subtexts, the following procedure was adopted. A dataset comprising 50 novels (see Tables 2 and 3) was used. Each book was split in subtexts comprising W tokens. If one considers a book as a sequence of tokens $\mathcal{W} = \{w_1, w_2, \dots\}$, the subtext \mathbb{T}_i will contain the sequence $\{w_{S_i}, w_{S_i+1}, \dots, w_{S_i+W}\}$, where $S_i = W \cdot i + 1$ and $i \in \mathbb{N}$. The variability of a given

measurement X across distinct subtexts \mathbb{T}_i 's of a given full book will be given by the coefficient of variation

$$\nu(X) = \left[\frac{\langle X^2 \rangle}{\langle X \rangle^2} - 1 \right]^{1/2}. \quad (8)$$

The variability of the following measurements were investigated in the current paper

$$X = \{\langle \alpha^{(h=2)} \rangle, \langle \alpha^{(h=3)} \rangle, \langle k_n \rangle, \langle B \rangle, \langle C \rangle, \quad (9)$$

$$\langle I \rangle, \Delta I, \gamma(I), \langle l \rangle, \Delta l \text{ and } \gamma(l)\}, \quad (10)$$

where $\langle \dots \rangle$, Δ and γ represent the mean, the standard deviation and the skewness of the distribution of the measurements in a given subtext. The accessibility was used to measure the prominence of a given word considering its nearest concentric neighborhood ($h = 2$ and $h = 3$). Higher values of h were not employed because the accessibility computed in higher levels is not informative [61]. The clustering coefficient and the average nearest neighbors degrees were used to quantify the connectivity between neighbors. The global prominence was measured with the average shortest path lengths and betweenness. Finally, the intermittency was employed to quantify the relevance of the words according to their spatial distribution along the text.

The variability obtained for each X across distinct subtexts is shown in Figure 2. The results confirm the the variability $\nu(X)$ of all measurements studied shows the same behavior, as revealed by a decreasing tendency as W increases. This means that the statistical fluctuations across distinct subtexts decrease as larger portions of subtexts are considered. The majority of the measurements displayed a variability below 0.35 for $W \geq 1,500$. The average accessibility displayed a typical coefficient of variation below 0.20 for $W \geq 1,500$. The average neighbor degree turned out to be the measurement taking the lowest values of variability. Even when very small pieces of texts were taken into account ($W=300$), the typical variability was always below 0,20. The average betweenness also took low values of variability for small subtexts. However, the lowest values were found for $W \geq 1,500$. The average clustering coefficient was one of the measurements whose variability across subtexts displayed a high dependence upon W . More specifically, for small subtexts, $\langle C \rangle$ turned out to be unstable, as revealed by coefficient of variations surpassing $\nu = 0.65$. This result shows that $\langle C \rangle$ should not be employed for the topological analysis of small texts because it is very sensitive to the sampling size. Concerning the intermittency, both $\langle I \rangle$ and ΔI displayed low values of variability for $W \geq 600$. Conversely, the skewness $\gamma(I)$ displayed high variabilities even for large texts ($W = 2,100$). This result might be a consequence of the fact that $\gamma(I)$ reflects the fraction of keywords in a text [57]. Therefore, if the amount of relevant words in each subtext presents a high variability, it is natural to expect that $\gamma(I)$ will vary accordingly. With regard to the shortest path lengths, both $\langle l \rangle$ and Δl displayed low values of fluctuations for $W \geq 1,500$. Similarly to $\gamma(I)$, $\gamma(l)$ presented a high dispersion across distinct subtexts.

All in all, the results displayed in Figure 2 reveal the most of measurements displays low statistical fluctuations when shorter texts are analyzed. The only exceptions were

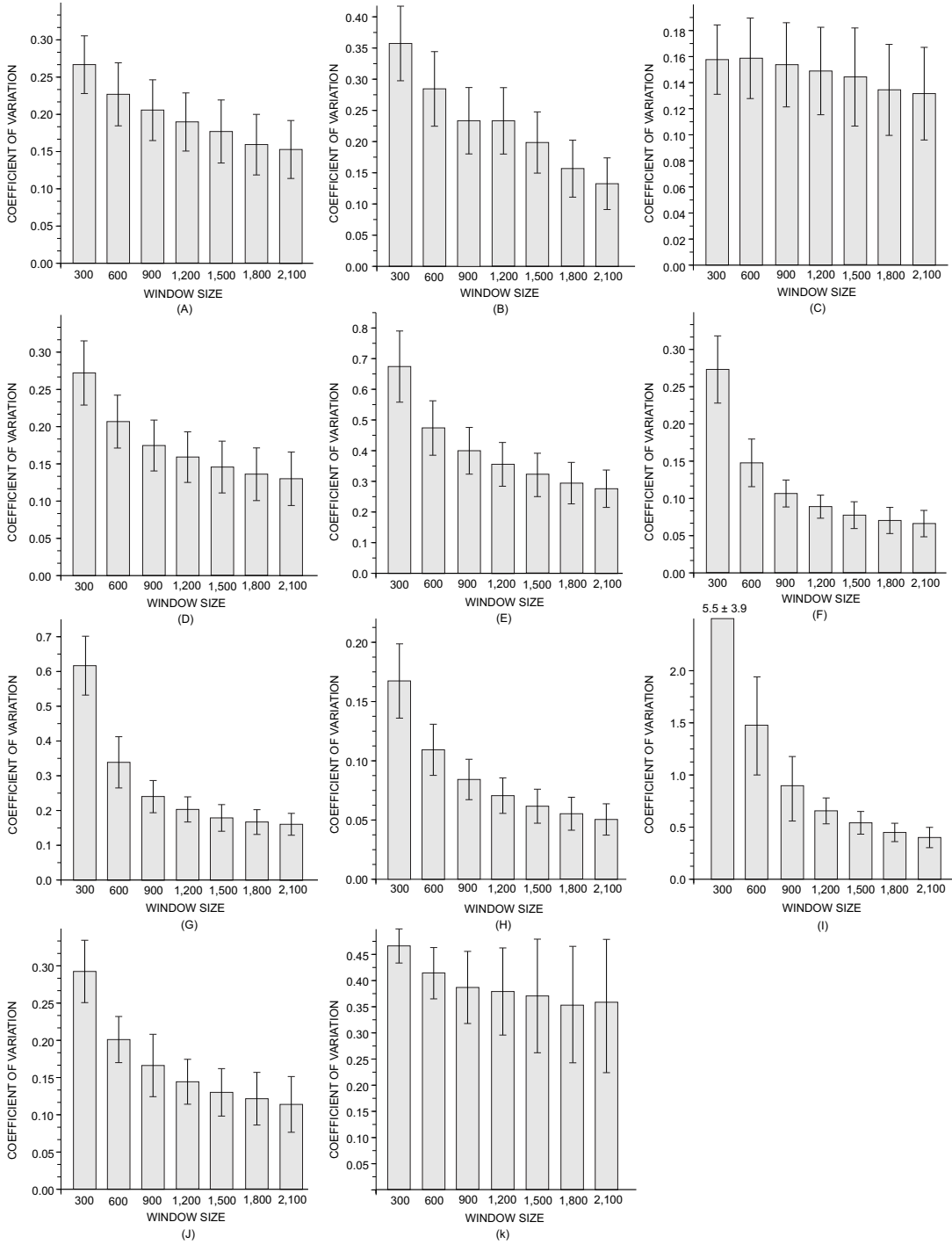


Figure 2. Coefficient of variations for the measurements (a) average accessibility $\langle \alpha^{(h=2)} \rangle$; (b) average accessibility $\langle \alpha^{(h=3)} \rangle$; (c) average neighbors degree $\langle k_n \rangle$; (d) average betweenness $\langle B \rangle$; (e) average clustering coefficient $\langle C \rangle$; (f) average intermittency $\langle I \rangle$; (g) standard deviation of the intermittency ΔI ; (h) skewness of the intermittency $\gamma(I)$; (i) average shortest path length $\langle l \rangle$; (j) standard deviation of the shortest path length Δl ; and (k) skewness of the intermittency $\gamma(I)$.

Table 2. List of Books (part 1).

Date	Author	Book
1811	Jane Austen	Sense and Sensibility
1815	Jane Austen	Emma
1826	James F. Cooper	The Last of the Mohicans
1841	Charles Dickens	Barnaby Rudge: A Tale of the Riots of Eighty
1842	Charles Darwin	The Structure and Distribution of Coral Reefs
1842	Charles Dickens	American Notes for General Circulation
1844	Charles Darwin	Geological Observations on the Volcanic Islands
1844	Charles Darwin	Geological Observations on the South America
1847	Charlotte Bronte	Jane Eyre
1847	William M. Thackeray	Vanity Fair: A Novel without a Hero
1847	Emily Bronte	Wuthering Heights
1850	Charles Dickens	David Copperfield
1851	Herman Melville	Moby-Dick; or, The Whale
1854	Charles Dickens	Hard Times – For These Times
1856	Gustave Flaubert	Madame Bovary
1859	Charles Dickens	A Tale of Two Cities
1859	Wilkie Collins	Woman in White
1861	Charles Dickens	Great Expectations
1868	Louisa May Alcott	Little Women
1869	Mark Twain	The Innocents Abroad
1869	Leo Tolstoy	War and Peace
1872	Charles Darwin	The Expression of the Emotions in Man and Animals
1873	Thomas Hardy	A Pair of Blue Eyes
1874	Thomas Hardy	Far From the Madding Crowd
1876	Thomas Hardy	The Hand of Ethelberta: A Comedy in Chapters

List of books employed for the analysis of variability of complex network measurements.

the skewness of the average shortest path lengths and the skewness of the intermittency. In both cases, the variability remained high even for subtexts comprising 2,100 tokens. In any case, in general, it is reasonable to suppose that a proper sampling allows a proper characterization of the *local* topological properties of books. In order to verify the applicability of sampling books in real stylometric tasks, the next section investigates how the sampling affects the performance of the authorship recognition task [71].

Authorship recognition via topological analysis of subtexts

Authorship recognition methods are important because they can be applied e.g to solve copyright disagreements [72], to intercept terrorist messages [73] and to classify

Table 3. List of Books (part 2).

Date	Author	Book
1876	George Eliot	Daniel Deronda
1877	Leo Tolstoy	Anna Karenina
1877	Charles Darwin	The Different Forms of Flowers on Plants of the Same Species
1883	Mark Twain	Life on the Mississippi
1884	Mark Twain	Adventures of Huckleberry Finn
1886	Thomas Hardy	The Mayor of Casterbridge
1887	Arthur Conan Doyle	A Study in Scarlet
1895	Thomas Hardy	Jude the Obscure
1897	Arthur Conan Doyle	Uncle Bernac
1900	Arthur Conan Doyle	War in South Africa
1903	Bram Stoker	The Jewel of Seven Stars
1905	Bram Stoker	The Man
1909	Bram Stoker	The Lady of the Shroud
1911	Bram Stoker	The Lair of the White Worm
1912	Arthur Conan Doyle	The Lost World
1914	Bram Stoker	Dracula's Guest
1914	Arthur Conan Doyle	The Valley of Fear
1915	P. G. Wodehouse	Something New
1915	Virginia Woolf	The Voyage Out
1920	Edith Wharton	The Age of Innocence
1921	P. G. Wodehouse	The Girl on the Boat
1921	P. G. Wodehouse	Indiscretions of Archie
1922	P. G. Wodehouse	The Adventures of Sally
1922	P. G. Wodehouse	The Clicking of Cuthbert

List of books employed for the analysis of variability of complex network measurements.

literary manuscripts [74]. Automatic authorship recognition techniques became popular after the famous investigation of Mosteller and Wallace on the Federalist Papers [75]. After this seminal study, researchers have tried to discover novel features to quantify styles, i.e. the textual properties that unequivocally identify authors. Currently, this line of research is known as stylometry [71]. Traditional features employed to discriminate authors include statistical properties of words (average length, frequency and intermittency of specific words, richness of vocabulary) [49] and characters (frequency and correlations) [72]. In addition to the lexical features, syntactical (frequency of specific parts-of-speech or chunks) and semantical properties (semantical dependencies) have been employed [49]. Current research have devised new attributes for the development of robust classifiers [49]. In recent years, it has been shown that the topological properties of complex networks has been useful to capture various

textual attributes related to authors' styles [20, 42–46]. For example, a significant discriminability of authors could be found in [57]. Because authors leave stylistic marks on the topological structure of complex networks [46, 57, 58], in this section the authorship recognition problem is studied by measuring the topological properties of co-occurrence networks. More specifically, the effects of the sampling on the classification are investigated.

For the classification task, the list of books and authors considered is shown in Table 4. Subtext lengths ranging in the interval $500 \leq W \leq 21,400$ were considered. To evaluate the performance of the task in short texts, note that low values of W were used. A comparison of the performance obtained with short and large texts was carried out by analyzing the accuracy rates found for higher values of W . In the dataset shown in Table 4, the larger window considered ($W = 21,400$) corresponds to the case where no sampling was performed. In other words, when $W = 21,400$ each book was represented by a single subtext. The values of $W = 7,130$ and $W = 5,350$ correspond to the division of full texts in three and four parts, respectively. Four classifiers were employed to perform the supervised classifications: (i) nearest neighbors, (ii) naive Bayes, (iii) decision trees, and (iv) support vector machines.

The performance obtained for the authorship recognition task is shown in Table 5. A visualization of the classification is provided in Figure 3. The highest accuracy rate found was 86.67% (p-value $< 10^{-10}$), which confirms that the topological characterization of subtexts is able to discriminate authors. In all classifiers, the lowest accuracy rates were found for the shortest subtexts ($W = 500$). This poor discriminability of authors can be attributed to the very short length employed for the task (see Figure 4). In order to compare the performance obtained with short and full texts, it is possible to define a threshold W_L from which the accuracy rate surpasses the value $\theta \times \text{AFB}$, where AFB is the accuracy found with the traditional approach based on full texts. Using $\theta = 0.85$, the following thresholds were obtained: $W_L = 1,600$ (kNN), $W_L = 1,700$ (Bayes), $W_L = 1,000$ (C4.5) and $W_L = 1,000$ (SVM). Therefore, in the dataset provided in Table 4, one can obtain 85% of the discriminability found with full books ($W = 21,400$) if one analyzes subtexts comprising at least $W = 1,700$ tokens.

If one compares the accuracy rates found for distinct values of W , it is interesting to note that the highest accuracy rates does not occur when full books are used. As a matter of fact, the highest accuracy rates for kNN, Bayes, C4.5 and SVM were obtained for $W = 21,400$, $W = 5,350$, $W = 7,130$ and $W = 5,350$, respectively. This observation suggests that, when a suitable sampling is performed, the performance of the classification can even be improved. In line with the results reported in the literature [76], the SVM outperformed other classifiers (for a given subtext length). In particular, this classifier yielded high accuracy rates even for short subtexts ($W = 2,500$). The accuracy obtained for the SVM when $W = 3,000$ even surpassed all the accuracies found for the other classifiers. This means that the combination of local characterizations via sampling with support vector machines outperformed the

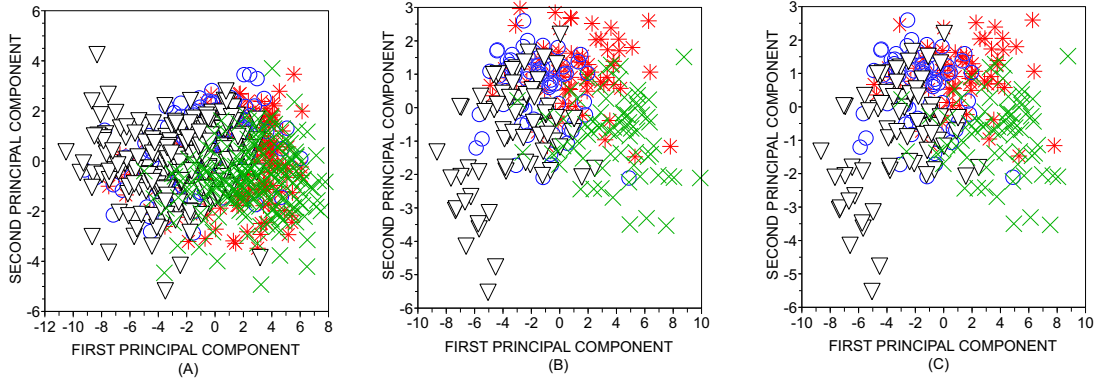


Figure 3. Visualization of the projection of the eleven attributes employed for the classification of authorship. The principal component analysis was employed. The length of the subtexts considered for the purpose of authorship recognition were: (a) $W = 500$; (b) $W = 1,000$; and (c) $W = 1,500$. Note that the discriminability increases as the subtexts become larger.

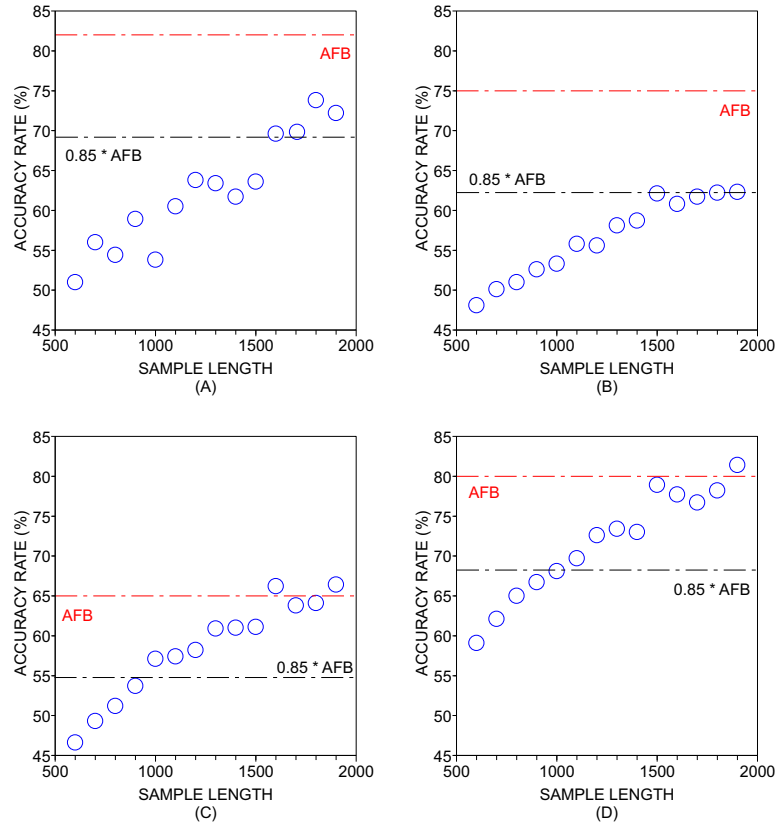


Figure 4. Accuracy rate as a function of the subtext length. The classifiers employed were (a) kNN; (b) Naive Bayes; (c) C4.5; and (d) SVM. The accuracy rate obtained with full texts (AFB) is represented as a red dashed line. Considering all the classifiers, one can obtain 85% of the discriminability found with full books ($W = 21,400$) if one analyzes subtexts comprising at least $W = 1,700$ tokens.

Table 4. List of books employed for the authorship recognitions task.

Date	Author	Book
1892	Arthur C. Doyle	The Adventures of Sherlock Holmes
1907	Arthur C. Doyle	Through the Magic Door
1898	Arthur C. Doyle	The Tragedy of Korosko
1915	Arthur C. Doyle	The Valley of Fear
1902	Arthur C. Doyle	The War in South Africa
1914	Bram Stoker	Dracula's Guest
1903	Bram Stoker	The Jewel of Seven Stars
1909	Bram Stoker	The Lady of the Shroud
1911	Bram Stoker	The Lair of the White Worm
1905	Bram Stoker	The Man
1842	Charles Darwin	The Structure and Distribution of Coral Reefs
1877	Charles Darwin	The Different Forms of Flowers
1872	Charles Darwin	The Expression of the Emotions in Man and Animals
1844	Charles Darwin	Geological Observations on the Volcanic Islands
1844	Charles Darwin	Geological Observations on the South America
1914	Hector H. Munro	Beasts and Super Beasts
1911	Hector H. Munro	The Chronicles of Clovis
1919	Hector H. Munro	Toys of Peace
1912	Hector H. Munro	The Unbearable Bassington
1913	Hector H. Munro	When William Came

List of books employed for the authorship recognitions task. The authorship of four authors were evaluated: Arthur C. Doyle, Bram Stoker, Charles Darwin and Hector H. Munro (Saki).

classification based on full books. Given these observations, it seems that support vector machines are more robust than other classifiers when shorter texts are analyzed.

Conclusions

This study probed the influence of sampling texts in the topological analysis of word adjacency networks. An individual analysis of variability of each network measurement revealed that most of them display a low variability across samples. The only exceptions were the skewness of the average shortest path length and the skewness of the intermittency, as revealed by high values of variability across samples even for larger subtexts. Taken together, these results evidence that short pieces of texts are suitable for network analysis, because the sampling process yields weakened statistical fluctuations for short texts.

The influence of the sampling process on a practical classification task was also investigated. Surprisingly, high accuracy rates could be found for texts comprising 1,700

Table 5. Accuracy rate for the authorship recognition task.

Sample length	kNN (%)	Bayes (%)	C4.5 (%)	SVM (%)
500	49.60	46.70	48.10	56.90
1,000	53.33	53.30	57.10	68.09
1,500	63.57	62.10	61.07	78.92
2,000	72.50	63.50	63.50	78.50
2,500	71.25	66.87	68.75	81.25
3,000	68.57	67.14	59.28	83.57
5,350	75.00	77.50	80.00	86.25
7,130	81.60	73.30	81.60	86.67
21,400	82.00	75.00	65.00	80.00

Variation of accuracy rate as a function of subtext length W . The best accuracy rate was found for the SVM, when the sample length $W = 7,130$ tokens.

tokens, which amounts to less than 8% of the length of a full book. The SVM classifier turned out to be the best classifier for the authorship recognition task based on short texts, as it outperformed the other three traditional classifiers. These results confirm that, when sufficiently large texts are generated, the sampling does not significantly affect the performance of the classification. Actually, the local characterization of texts might even improve the performance of the classifiers. In addition to allowing the use of short texts in classification tasks based on stylometry via topological characterization of word adjacency networks, the use of small pieces of texts tends to reduce the effects of the so-called curse-of-dimensionality [65], as more training examples are included in the attribute space. A possible weakness of the sampling method is that it can only be applied to large texts. The sampling of short documents generates texts with high topological variability. In these cases, other models should be used to capture relevant features for the specific task [77]. The analysis of short texts has been performed, for example, to detect sentiments. When short comments in product reviews or tweets are analyzed, semantical methods methods have been applied [78, 79].

This paper showed, as a proof of principle, that smaller pieces of texts can also be useful in textual network analysis. Following this research line, future works could, for example, apply the techniques described here to identify stylistic inconsistencies in written texts. Such inconsistencies could be found, for example, by identifying topological outliers, i.e. subtexts whose topology is different from other observations in the same book. The identification of stylistic inconsistencies could also be useful for recognizing multiple authorship or even plagiarisms [80]. Most importantly, the techniques described here could also be extended in a straightforward fashion to study written texts as temporal series [81], thus allowing the study of texts as time-varying complex networks [82].

Acknowledgments

I acknowledge financial support from São Paulo Research Foundation (FAPESP-Brazil) (grant number 13/06717-4).

References

- [1] Newman M (2010) Networks: an introduction. Oxford University Press, Inc.
- [2] Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393 (6684) : 440–442.
- [3] Dangalchev C. (2004) Generation models for scale-free networks. *Physica A* 338 : 659–671.
- [4] Doye JPK, Massen CP (2005) Characterizing the network topology of the energy landscapes of atomic clusters. *The Journal of Chemical Physics* 122 (8) : 084105.
- [5] Hughes D, Paczuski M (2004) A heavenly example of scale free networks and self-organized criticality. *Physica A* 342 (1–2) : 158–163.
- [6] Girvan M, Newman MEJ (2002) Community structure in social and biological networks. *Proceedings of the National Academy of Science USA* 99 (12) : 7821–7826.
- [7] Freeman LC (1979) Centrality in social networks: conceptual clarification. *Social Networks* 1 : 215–239.
- [8] Toivonen R, Kovanen L, Kivela M, Onnela J-P, Saramaki J, Kaski K (2009) A comparative study of social network models: network evolution models and nodal attribute models. *Social Networks* 31 (4): 240–254.
- [9] Barabási A-L and Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* 5 (2) : 101–113.
- [10] Betel D, Isserlin R, Hogue CWV (2004). Analysis of domain correlations in yeast protein complexes. *Bioinformatics* 20 55–62.
- [11] Bullmore ET, Sporns O (2009) Complex brain networks: graph-theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience* 10 186–198.
- [12] Rubinov M, Sporns O (2010) Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52 1059–1069.
- [13] Stam CJ, Reijneveld JC (2007) Graph theoretical analysis of complex networks in the brain. *Nonlinear Biomedical Physics* 1:3.
- [14] Baronchelli AA, Ferrer-i-Cancho R, Pastor-Satorras R, Chater N and Christiansen MH (2013) Networks in cognitive sciences. *Trends in Cognitive Science* 17(7): 348–360.
- [15] Cano P, Celma O, Koppenberger M, Buld JM (2006) Topology of music recommendation networks. *Chaos* 16: 013107.
- [16] Backes AR, Casanova D, Bruno OM (2009) A complex network-based approach for boundary shape analysis. *Pattern Recognition* 42 (1): 54–67.
- [17] Moura APS, Lai YC, Motter AE (2003) Signatures of small-world and scale-free properties in large computer programs. *Physical Review E* 68 (1): 017102.
- [18] Silva TC, Amancio DR (2012) Word sense disambiguation via high order of learning in complex networks. *Europhysics Letters* 98 58001.
- [19] Kong JS, Rezaei BA, Sarshar N, Roychowdhury VP, Boykin PO (2006) Collaborative spam filtering using e-mail networks. *Computer* 39 (8): 67–73.
- [20] Cong J, Liu H (2014) Approaching human language with complex networks. *Physics of Life Reviews* (to appear) DOI: 10.1016/j.plrev.2014.04.004.
- [21] Liu H (2008) The complexity of Chinese syntactic dependency networks. *Physica A* 387 (12): 3048–3058.
- [22] Liu H, Li W (2010) Language clusters based on linguistic complex networks. *Chinese Science Bulletin* 55 (30): 3458–3465.
- [23] Abramov O, Mehler A (2011) Automatic language classification by means of syntactic dependency networks. *Journal of Quantitative Linguistics* 18 (4): 291–336.
- [24] Sporns O, Chialvo DR, Kaiser M, Hilgetag CC (2004) Organization, development and function of complex brain networks. *Trends in Cognitive Sciences* 8 (9): 418–425.
- [25] Zamora-Lopez G, Russo E, Gleiser P M, Zhou CS, Kurths J (2011) Characterizing the complexity of brain and mind networks. *Philosophical Transactions of the Royal Society A* 369, 3730–3747.

- [26] Arruda GF, Costa LF, Schubert D, Rodrigues FA (2013) Structure and dynamics of functional networks in child-onset schizophrenia. *Clinical Neurophysiology*, 125(8): 1589–1595.
- [27] Barttfeld P, Wicker B, Cukier S, Navarta S, Lew S, Sigman M (2011) A big-world network in ASD: dynamical connectivity analysis reflects a deficit in long-range connections and an excess of short-range connections. *Neuropsychologia* 49, 254–263.
- [28] Zhao X, Liu Y, Wang X, Liu B, Xi Q, Guo Q, Jiang H, Jiang T, Wang P (2012) Disrupted small-world brain networks in moderate Alzheimer’s disease: a resting-state fMRI study. *PLoS One* 7, e33540.
- [29] Borge-Holthoefer J, Arenas A (2010) Semantic Networks: Structure and Dynamics. *Entropy*, 12(5) 1264–1302.
- [30] Beckage N, Smith L, Hills T (2011) Small worlds and semantic network growth in typical and late talkers. *PLoS One* 6, e19348.
- [31] Vitevitch MS, Ercal G, Adagarla B (2011) Simulating retrieval from a highly clustered network: implications for spoken word recognition. *Frontiers in Psychology* 2, 369.
- [32] Amancio DR, Oliveira ON, Costa LF (2011) On the concepts of complex networks to quantify the difficulty in finding the way out of labyrinths. *Physica A* 390, 4673–4683.
- [33] Ferrer i Cancho R, Solé R, Kohler R (2004) Patterns in syntactic dependency networks. *Physical Review E* 69 51915.
- [34] Amancio DR, Nunes MG, Oliveira Jr. ON, Costa LdF (2012) Extractive summarization using complex networks and syntactic dependency. *Physica A* 391 1855–1864.
- [35] Liu H. Statistical properties of Chinese semantic networks. *Chinese Science Bulletin* 2009; 54: 2781–2785.
- [36] Masucci AP, Kalampokis A, Eguluz VM, Hernandez-Garcia E (2011) Wikipedia information flow analysis reveals the scale-free architecture of the semantic space. *PLoS ONE* 6 e17333.
- [37] Silva TC, Amancio DR (2013) Discriminating word senses with tourist walks in complex networks. *The European Physical Journal B* 86: 297.
- [38] Matveeva I, Levow G-A (2006) Graph-based generalized Latent Semantic Analysis for document representation. *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, 61–64.
- [39] Landauer TK, Laham D, Derr M (2004) From paragraph to graph: latent semantic analysis for information visualization. *Proceedings of the National Academy of Science* 101, 5214–5219.
- [40] Henderson K, Eliassi-Rad T (2009) Applying Latent Dirichlet Allocation to group discovery in large graphs. *Proceedings of the ACM symposium on Applied Computing*, ACM, New York, NY, USA, 1456–1461.
- [41] Amancio DR, Altmann EG, Rybski D, Oliveira Jr. ON, Costa LdF (2013) Probing the statistical properties of unknown texts: application to the Voynich manuscript. *PLoS ONE* 8 e67310.
- [42] Amancio DR, Aluisio SM, Oliveira ON, Costa LdF (2012) Complex networks analysis of language complexity. *Europhysics Letters* 100 58002.
- [43] Amancio DR, Oliveira Jr. ON, Costa LdF (2012) Identification of literary movements using complex networks to represent texts. *New Journal of Physics* 14 043029.
- [44] Roxas-Villanueva RM, Nambatac MK, Tapang G (2012) Characterizing english poetic style using complex networks. *International Journal of Modern Physics C* 23 1250009.
- [45] Grabska-Gradzinska I, Kulig A, Kwapien J, Drozd S (2012) Complex network analysis of literary and scientific texts. *International Journal of Modern Physics C* 23 1250051.
- [46] Roxas RM, Tapang G (2010) Prose and poetry classification and boundary detection using word adjacency network analysis. *International Journal of Modern Physics C* 21 503.
- [47] Carron PM, Kenna R (2013) Network analysis of the *Islendinga sogur* – the Sagas of Icelanders. *The European Physical Journal B* 86 407.
- [48] Liu HT, Cong J (2013) Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin* 58 (10): 1139–1144.

- [49] Stamatatos E (2009) A survey of modern authorship attribution methods (2009) *Journal of the Association for Information Science and Technology* 60, 538–556.
- [50] Mihalcea R, Radev D (2011) *Graph-based natural language processing and information retrieval*. Cambridge University Press.
- [51] Navigli R (2009) Word sense disambiguation: a survey. *ACM Computing Surveys* 41 (2): 1–69.
- [52] Dunning T (1993) Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.* 19 (1) 61–74.
- [53] Ferrer-i-Cancho R, Sole RV (2001) The small-world of human language. *Proc. R. Soc. Lond. B* 268 (1482) 2261–2265.
- [54] Masucci AP, Rodgers GJ (2006) Network properties of written human language. *Physical Review E* 74 (2), 026102.
- [55] Veronis J (2004) HyperLex: lexical cartography for information retrieval. *Computer Speech and Language* 18(3): 223–252.
- [56] Lin D (1998) Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2 (ACL '98)* Stroudsburg, PA, USA, 768–774
- [57] Amancio DR, Altmann EG, Oliveira Jr ON, Costa LdF (2011) Comparing intermittency and network measurements of words and their dependence on authorship. *New Journal of Physics* 13 123024.
- [58] Antiqueira L, Pardo TAS, Nunes MGV, Oliveira Jr. ON (2007) Some issues on complex networks for author characterization. *Inteligencia Artificial* 11, 51-58.
- [59] Berger AL, Della Pietra SA, Della Pietra VJ (1996) A maximum entropy approach to natural language processing. *Computational Linguistics* 22 39–71.
- [60] Amancio DR, Oliveira Jr. ON, Costa LF (2012) Structure semantics interplay in complex networks and its effects on the predictability of similarity in texts. *Physica A* 391, 4406–4419.
- [61] Costa LF, Silva FN (2006) Hierarchical characterization of complex networks. *Journal of Statistical Physics* 215(4): 841-872.
- [62] Carretero-Campos C, Bernaola-Galvn P, Coronado AV, Carpena P (2013) Improving statistical keyword detection in short texts: entropic and clustering approaches. *Physica A* 392: 1481–1492.
- [63] Freeman L (1977) A set of measures of centrality based on betweenness. *Sociometry* 40: 3541.
- [64] Herrera JP, Pury PA (2008) Statistical keyword detection in literary corpora. *The European Physical Journal B* 63: 135-146.
- [65] Duda RO, Hart PE, Stork DG (2000) *Pattern Classification*. Wiley-Interscience.
- [66] Darrell T, Indyk P, Shakhnarovich F (2006). *Nearest neighbor methods in learning and vision: theory and practice*. MIT Press.
- [67] Murthy SK (1998) Automatic construction of decision trees from data: a multi-disciplinary survey. *Data Min. Knowl. Discov* 2: 345–389.
- [68] Hand DJ, Yu K (2001) Idiot’s Bayes: not so stupid after all? *Statistical Review* 69: 385–398.
- [69] Cortes C, Vapnik V (1995) Support-Vector networks. *Machine Learning* 20: 273.
- [70] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D (2007) Top 10 algorithms in data mining. *Knowl. Inf. Syst.* 14 (1): 1–37.
- [71] Juola P (2006) Authorship attribution. *Foundations and Trends in Information Retrieval* 1 : 3.
- [72] Grant TD (2007) Quantifying evidence for forensic authorship analysis. *International Journal of Speech Language and the Law* 14(1): 1–25.
- [73] Abbasi A, Chen H (2005) Applying authorship analysis to extremist-group Web forum messages, *IEEE Intelligent Systems* 20 (5): 67–75.
- [74] Ebrahimpour M, Putnins TJ, Berryman MJ, Allison A, Ng BW-H, Derek A (2013) Automated authorship attribution using advanced signal classification techniques. *PLoS ONE* 8(2): e54998.
- [75] Mosteller F, Wallace DL (1964) *Inference and disputed authorship: The Federalist*. Addison-

Wesley.

- [76] Amancio DR, Comin CH, Casanova D, Travieso G, Bruno OM, Rodrigues FA, Costa LdF (2014) A systematic comparison of supervised classifiers. PLoS ONE 9 e94137.
- [77] Manning CD and Schutze H (1999) Foundations of statistical natural language processing. MIT Press, Cambridge, MA, USA.
- [78] Golder SA, Macy MW (2011) Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science 333 (6051):1878–81
- [79] Cambria E, Schuller B, Xia Y, Havasi C (2013) New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems 28(2): 15–21.
- [80] Dreher H (2007) Automatic conceptual analysis for plagiarism detection. Information and Beyond: The Journal of Issues in Informing Science and Information Technology 4: 601–614.
- [81] Amancio DR (2014) Patterns in stylistic fluctuations of written texts. Manuscript under review.
- [82] Karsai M, Perra N, Vespignani A (2014) Time-varying networks and the weakness of strong ties. Scientific Reports 4 4001.